Visual Analytics (INM433)

Daniel Dixey

MSc Data Science

Wednesday 8th April

Data

Overview, description and properties of the dataset

Data [10%]. Description of the data chosen for the analysis: type, structure, size, properties of the components.

CrossFit is a rigorous fitness methodology that that attempts to unite multiple domains of fitness into one programme; weightlifting, High Intensity Interval training and gymnastics, where the primary aim is to prepare all participants to be ready for any physical challenge that may arise. The dataset that has been obtained and used in this report has been taken from the CrossFit Open leader board on the CrossFit Games website [LINK]. The Open is a worldwide competition that aims to bring to the community of CrossFit together via a series of five weekly physical workouts.

Class of Data - Object Referenced Data
Completeness – For all the attributes there exists a degree of incompleteness, during the analysis missing values will not be inferred or interpolated. They will be ignored at the expense of introducing bias results, however they will be a true representation of the sample of data that is available.
Uncertainty – The data that has been collected is based on the values that have been supplied by the individuals who have registered. The credibility and integrity of the data could be of some concern, where necessary spurious and obviously false (age >100 years and weight <20kg) information will be removed where deemed necessary during the analysis.

The dataset has been acquired using a web scraping technique, the details of the methodology of this scraping have not been disclosed. Typically, the most widely adopted method of extracting the data is through the use of a web crawling application. A web crawler, it would navigate a website recursively according to a set of defined user parameters to extract the underlying web script of each page. Once the collection has been completed the data is parsed so that the desired features, in this case the features of the dataset, are obtained and output to a tabular format, in this case a csv format, which is then ready for further processing and eventually exploratory data analysis.

Analysis Tasks

- Literature review, work to be undertaken and synoptic task

Analysis tasks [10%]. Analysis task(s) chosen for your analysis:
Do these relate to the tasks you addressed in your literature review (part 1)? If so, how?
Specific task formulations for the chosen data and the corresponding generic task types.

Dimensionality Reduction

As part of the Literature Review, dimensionality reduction was discussed. The focus was deliberately limited to this area as the number of considerations and techniques available in the area of multidimensional analysis are vast. Although DR is not the only technique of analysing multidimensional data available, dimensionality reduction was specifically identified as very useful in assisting with the visualisation of multidimensional datasets. When a dataset is of a large magnitude of dimensions the Analyst who is trying to understand the data model would find it difficult to interpret and make meaningful interpretations of the data. For this reasoning, by example, a technique called Principal Component Analysis (PCA) enables you to reduce a number of degrees of freedom without the loss of information. No loss of information is achieved as the eigenvalues that are generated multiplied by the normalised data will return the original dataset, this a valuable property as it means any interesting patterns will not be lost as a result of implementing this method.

Tasks of the Analysis:

- How gender, height and weight relate to the performance of athletes in the Crossfit Open competition?"
  - Who are the top individuals globally and by region? Are they comparative and exhibit the same types of characteristics?
  - How does the drop of participation of each of the Open events distort the comparison of athletes.
  - What key traits do the top athletes have that the less competitive do not?

Analysis Methodology

- Overview, description and techniques used in the process

Analysis methodology [30%]. Methodology of your analysis.
Does this relate to those in your literature review (part 1) and/or a lecture/practical? If do, how? Did you add any modifcations? If so, what and why?
Description of your methodology.
How the analytical labour is divided between the human and computer.
Computational methods involved, types of their inputs and outputs.
Data transformations (if any).
Visual and interactive techniques involved. How the visualisation supports the human reasoning.

Description of Methodology

3.2.1 Preprocessing and Cleaning

As a result of the web crawling the data is messy and the data will require cleaning to ensure that the data is in usable format. The types of issues that are prevalent throughout the dataset are; missing values/data, the units (kg/lb/cm) used for some variables are inconsistent (erroneous values) and finally the names will require delimitation to get the names in a usable format.

3.2.2 Transformations

Unsupervised learning techniques like Gaussian Mixed Methods (GMM) will be used to understand and gain insight statistically as to the number of clusters that exist within a dataset. The output of this method will generate additional dimensions with labels corresponding the the group and probability of belonging to that group. This method of clustering offers a more robust (statistical) use over k-means clustering, the reasoning for this is that domain knowledge could inhibit the information that could be gained for not using it however it can also be used to confirm domain specific knowledge. My own perceived interpretation of the data will should not influence the decision of the number of athlete groups in the dataset and this is an advantage of using GMM over k-means. The second expected transformation is the calculation is of a distance matrix. This will be used for determining similarities between dimensions in the dataset and also for the use in projections, specifically the Igloo style plots (semi-circle projection).

3.3 Division of analytical labour

A five phase breakdown of how both the division of human and computational work will be adopted is described below:

Understanding of the characteristics of the attributes

The first stage of the analysis is to get an understanding of each of the distributions and basics statistics about every attribute. This will be achieved through the use of histograms, this type of graphic offers the most accessible means of understanding the distribution of the data. When a histogram is used in conjunction with a table of basic metrics about the attribute it will be possible to interpret the magnitude of the values in the attribute and support human reasoning. Information that I will be expected to know at the end of this phase are; the type of distribution (Gaussian, Poisson), do outliers exist and limitations of the attribute (missing values, can or cannot normalisation be applied if require in a later stage) and finally if the attribute requires a transformation. With the information I will be able to able data mining and visualisation techniques according based on the restrictions of each attribute.

Understanding of the Relationships between the attributes

It is initially though that this stage will involve looking at a combination of histograms and scatter plots to understand the relationships between attributes. This will be achieved computation through the use of visual encodings: shape, size, colour and labelling, to aid with the understanding of one or many attributes are related and correlated to one another. The use of different visual encodings will be paramount at this stage which are.....

Describing the relationships

At this stage a good understanding of the the attributes will be expected to be known. This stage will involve explaining the relationships that have been found in phases one and two. Two computational methods will be utilised at this stage; Gaussian Mixture Models (GMM) and Self Organising Maps (SOM), both methods deal well with high dimensionality in datasets. EXPLAIN. Prior knowledges expects that there will be at least three distinct groups however the methods chosen do not require you to enter a kernal number in as one of the parameters unlike k-means clustering. One of the main advantages of using the two methods over k-means was described in an earlier section.

As the use of a Autoencoder was used in one of the papers review in the Literature Review, a number of models will be generated to see if there can be benefit for the dataset to identify interesting patterns. The advantage of this type of method is that if a non-linear relationship exist between features then in theory the Autoencoder would be able to identify this relationship and “learn” them. The reason this is possible is due to the type of activation function used in each layer, if for example a hyperbolic tangent function [] is used then it should be possible for the ANN to learn it. Due to the nature of ANN it is difficult to determine the most appropriate initial starting conditions for a network, finding “optimised” parameters requires time and so therefore will be a big drawback of this type of method.

Exploring the findings iteratively through interaction

It is expected that navigating many times between phase 1-3, the reasoning for this is that it is likely that the data and analysis will provoke more questions about the data. The canvas framework nature of Python Orange support this type of progressive investigation well with the use of the work flows, figure X.

Final presentation

Improvement of the visualisations will be implement here, this will involve using a number of Modules other than Orange for refinement and to effectively apply visual encodings.

Implementation

- Literature review, work to be undertaken and synoptic task

Implementation [30%]. Implementation of the analysis methodology: software used, links between methods (integrated in the same software or data transfer).

All of the initial preprocessing tasks of the dataset that are mentioned in Section 3.1.2 will be undertaken in Libre Office Calc (Linux Excel equivalent). The argument for the use of Calc is to recover the most amount of data possible, as opposed to applying generic rules in a scripting language like Python. The advantage of doing it this way is that if there are any expected or erroneous values then these can be accounted for at a more granular level. For the second phase of the implementation the data is required to be in a CSV format such that it can be read into the module.

Once this initial phase has been completed the data will be loaded in Python, for the majority of the exploratory analysis and visualisation a module called Orange (python-orange) will be utilised. Orange is data mining module that supports visualisation and analysis through a user interface called canvas. Orange Canvas as shown in figure **X** supports the use of visual work flows that enable you to keep track of your analysis and exploratory journey through the dataset. There are a number of benefits with this type of interface is that it allows that it cleverly aids with building and testing the users various ideas and visualisations, while also keeping track through the use of the widgets and work flows. By design this also supports the iterative nature of visual analytics which is where the use of computational and visual methods are maximised in potential in order to extract knowledge from the data. The main advantage over a predominately scripting approach is that many more visuals and variations of data mining techniques can be applied in a much shorter time frame. This advantage is also a shortfall of this type of approach as it is restricted by the limited visual options that the module offers.

The expectation is that all the work will be conducted in Orange Canvas, as the number of widgets available is vast. The widgets have been categorised into nine types of process, the most relevant to this analysis are; Data, Visualize, Classify, Regression and Unsupervised. Orange Canvas supports the use of Machine Learning, both unsupervised and supervised, algorithms as well the more traditional data mining techniques like Principal Component Analysis (PCA). These methods are imported from the Scikit-Learn module in Python. Utilising this module is another advantage of Orange Canvas as if a concise understanding of how a method works is required then I can refer directly to the well documented Scikit Learn documentation. If for any reason Orange Canvas does not support the a particular method well then a scripting approach will be used as a work around however this will still be undertaken in Python.

Analysis process

- Literature review, work to be undertaken and synoptic task

Analysis process [10%]. Illustrated description of the analysis process, including all steps and intermediate results. The description must demonstrate fulfilment of the requirements set in section 1. The illustrations must include commented screenshots of the visual displays used in the analysis. The comments must explain what the displays show and how this information was used in the following analysis steps or contributes to the final result.

Display initial results

Recite all of the lecture notes here

Results and conclusion

- Current status of the report/analysis

Results and conclusion [10%]. To what extent the posed task(s) have been fulfilled?